Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 496282, 11 pages
http://dx.doi.org/10.1155/2013/496282

Research Article

Learning Rates for l^1-Regularized Kernel Classifiers

Hongzhi Tong,[1] Di-Rong Chen,[2] and Fenghong Yang[3]

[1] *School of Statistics, University of International Business and Economics, Beijing 100029, China*
[2] *Department of Mathematics and LMIB, Beijing University of Aeronautics and Astronautics, Beijing 100083, China*
[3] *School of Applied Mathematics, Central University of Finance and Economics, Beijing 100081, China*

Correspondence should be addressed to Hongzhi Tong; tonghz@uibe.edu.cn

Received 25 July 2013; Accepted 6 October 2013

Academic Editor: Huijun Gao

We consider a family of classification algorithms generated from a regularization kernel scheme associated with l^1-regularizer and convex loss function. Our main purpose is to provide an explicit convergence rate for the excess misclassification error of the produced classifiers. The error decomposition includes approximation error, hypothesis error, and sample error. We apply some novel techniques to estimate the hypothesis error and sample error. Learning rates are eventually derived under some assumptions on the kernel, the input space, the marginal distribution, and the approximation error.

1. Introduction

Let X be a compact subset of $\mathbb{R}^n$, $Y = \{-1, 1\}$. Classification algorithms produce binary classifiers $\mathscr{C} : X \to Y$, such a classifier $\mathscr{C}$ labels a class $\mathscr{C}(x) \in Y$ for each point $x \in X$. The prediction power of the classifier $\mathscr{C}$ is measured by its misclassification error. If ρ is a probability distribution on $Z := X \times Y$, then the misclassification error of $\mathscr{C}$ is defined by

$$\mathscr{R}(\mathscr{C}) := \text{Prob}\{\mathscr{C}(x) \neq y\} = \int_X \rho(y \neq \mathscr{C}(x) \mid x)\, d\rho_X. \quad (1)$$

Here ρ_X is the marginal distribution on X and $\rho(\cdot \mid x)$ is the conditional probability measure at x induced by ρ. The classifier minimizing the misclassification error is called the Bayes rule f_c and is given by

$$f_c(x) = \begin{cases} 1, & \text{if } \rho(y = 1 \mid x) \geq \rho(y = -1 \mid x), \\ -1, & \text{otherwise.} \end{cases} \quad (2)$$

The classifiers considered in this paper have the form $\text{sgn}(f)$, defined as $\text{sgn}(f)(x) = 1$, if $f(x) \geq 0$, and $\text{sgn}(f)(x) = -1$, if $f(x) < 0$, induced by real-valued functions $f : X \to \mathbb{R}$. Those functions are generated from a regularization scheme associated with convex loss function (see [1]).

Definition 1. A continuous function $V: \mathbb{R} \to \mathbb{R}^+$ is called a classifying loss (function) if it is convex, differentiable at 0 with $V'(0) < 0$, and 1 is the smallest real for which the value of V is zero.

Typical examples of classifying loss include

(1) hinge loss $V_h(t) = (1 - t)_+ = \max\{1 - t, 0\}$ for the classical support vector machines (SVM) classifier; see [2–5];

(2) least square loss $V_{\text{ls}}(t) = (1-t)^2$; see for example [6, 7];

(3) q-norm ($q > 1$) SVM loss $V_q(t) = (1 - t)_+^q$; see [8, 9].

The following concept describes the increment of V.

Definition 2. One says that V has a increment exponent $\theta \geq 1$, if there exists some $c_\theta > 0$ such that

$$|V(t)| \leq c_\theta (1 + |t|)^\theta, \qquad \left|V'_\pm(t)\right| \leq c_\theta (1 + |t|)^{\theta-1}, \quad \forall t \in \mathbb{R}, \quad (3)$$

where $V'_\pm$ denotes the right and left derivatives of V.

It is easy to see that V_h, V_{ls}, and V_q satisfy Definition 2 with increment exponent 1, 2, and q, respectively.

Let $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ be a set of samples independently drawn according to ρ; we call

$$\mathscr{E}_{\mathbf{z}}(f) := \mathscr{E}_{\mathbf{z}}^V(f) = \frac{1}{m}\sum_{i=1}^m V(y_i f(x_i)) \tag{4}$$

the empirical error with respect to $\mathbf{z}$. Regularized learning schemes are implemented by minimizing a penalized version of empirical error over a set of functions, called a hypothesis space.

Definition 3. Given a classifying loss V and a hypothesis space $\mathscr{H}$, $\Omega : \mathscr{H} \to \mathbb{R}^+$ is a penalty functional called regularizer that reflects the constrains imposed on functions from $\mathscr{H}$. The regularized classifier is then defined as $\operatorname{sgn}(f_{\mathbf{z}})$, where $f_{\mathbf{z}}$ is a minimizer of the following regularization scheme:

$$f_{\mathbf{z}} := \arg\min_{f\in\mathscr{H}} \{\mathscr{E}_{\mathbf{z}}(f) + \lambda\Omega(f)\}. \tag{5}$$

Here λ is a regularization parameter which may depend on the sample size $\lambda = \lambda(m)$ with $\lim_{m\to\infty}\lambda(m) = 0$.

Choosing different hypothesis spaces and regularizers in (5) will lead to different regularization algorithms. These learning algorithms are often based on a kernel function $K : X \times X \to \mathbb{R}$ (see, e.g., [10]). One way appears naturally when K is a Mercer kernel. Such a kernel is continuous, symmetric, and positive semidefinite on $X \times X$. The reproducing kernel Hilbert spaces (RKHS) $\mathscr{H}_K$ associated with the Mercer kernel K is defined [11] to be the completion of the linear span of functions $\{K_x := K(x, \cdot) : x \in X\}$ with the inner product:

$$\langle K_x, K_u\rangle_K = K(x, u), \tag{6}$$

and the reproducing property is given by

$$\langle K_x, f\rangle_K = f(x), \quad \forall x \in X,\ f \in \mathscr{H}_K. \tag{7}$$

By setting $\mathscr{H} = \mathscr{H}_K$, $\Omega(f) = \|f\|_K^2$, (5) becomes the classical regularized classification scheme:

$$\widetilde{f}_{\mathbf{z},\lambda} := \arg\min_{f\in\mathscr{H}_K} \left\{\mathscr{E}_{\mathbf{z}}(f) + \lambda\|f\|_K^2\right\}. \tag{8}$$

Its mathematical analysis has been well understood with various techniques in extensive literature (see, e.g., [4, 5, 8, 12–15]). In this paper we will consider a different regularization scheme in RKHS for classification; in our setting, the regularizer is l^1-norm of the coefficients in the kernel expansions over the sample points.

Definition 4. Let

$$\begin{aligned} \mathscr{H}_{K,\mathbf{z}} &:= \left\{\sum_{i=1}^m a_i K_{x_i} : a_i \in \mathbb{R}\right\}, \\ \Omega_{\mathbf{z}}(f) &:= \inf\left\{\sum_{i=1}^m |a_i| : f = \sum_{i=1}^m a_i K_{x_i}\right\}. \end{aligned} \tag{9}$$

Then the l^1-regularized classification scheme is given as

$$f_{\mathbf{z},\lambda} := \arg\min_{f\in\mathscr{H}_{K,\mathbf{z}}} \{\mathscr{E}_{\mathbf{z}}(f) + \lambda\Omega_{\mathbf{z}}(f)\}. \tag{10}$$

Algorithm (10) can be efficiently computed because it reduces to solve a convex optimization problem in a finite dimensional space $\mathscr{H}_{K,\mathbf{z}}$, containing linear combinations of kernels centered on the training points.

In the last ten years, learning with l^1-regularization has attracted much attention. The increasing interest is mainly brought by the progress of the Lasso algorithm [16–18] and compressive sensing [19, 20], in which l^1-regularizer is able to yield sparse representation of the resulting minimizer. Kernel methods formulate learning and estimation problems in RKHS of functions expanded in terms of kernels. There have been a series of papers to investigate the learning ability of coefficient-based regularization kernel regression methods (see, e.g., [21–25]). However, as we know, there are currently a few results on classification based on kernel designing. For example, [26] studies classification problem with hinge loss V_h and l^1 complexity regularization in a finite-dimensional hypothesis space spanned by a set of base functions. While it does not assume a kernel setting nor is it assumed that the expansion must be in terms of the sample points, so the problem of data-dependent hypothesis space is not present there. Although [27] provided an error analysis for linear programming SVM classifiers by means of a stepping-stone from quadratic programming SVM to linear programming SVM, no evidence shows that this method can still work for other classifying losses.

In this paper we will present an elaborate error analysis for algorithm (10), and we use a modified error decomposition technique that was firstly introduced in [28], by dealing with the approximation error, the hypothesis error, and the sample error, and we derive an explicit learning rate for classification scheme (10) under some assumptions.

2. Preliminaries

For a classifying loss V, we define the generalization error of $f : X \to \mathbb{R}$ as

$$\mathscr{E}(f) := \mathscr{E}^V(f) = \int_Z V(yf(x))\, d\rho. \tag{11}$$

Let f_ρ^V be a measurable function minimizing the generalization error:

$$f_\rho^V := \arg\min \mathscr{E}(f), \tag{12}$$

where the minimum is taken over all measurable functions. According to Theorem 3(c) in [12], we may always choose an f_ρ^V satisfying $f_\rho^V(x) \in [-1, 1]$ for each $x \in X$. This choice will be taken throughout the paper.

Estimating the excess misclassification error

$$\mathscr{R}(\operatorname{sgn}(f_{\mathbf{z},\lambda})) - \mathscr{R}(f_c) \tag{13}$$

for classification scheme (10) is our main purpose. The following comparison theorem (see [7, 8, 12]) describes the relationship between excess misclassification error and excess generalization error.

Proposition 5. *If V is a classifying loss, then, for any measurable function f,*

$$\mathscr{R}(\operatorname{sgn}(f)) - \mathscr{R}(f_c) \le \begin{cases} \mathscr{E}(f) - \mathscr{E}(f_c) & \text{if } V(t) = (1-t)_+, \\ c_V \sqrt{\mathscr{E}(f) - \mathscr{E}\left(f_\rho^V\right)} & \text{if } V''(0) \ge 0, \end{cases} \tag{14}$$

where c_V is some constant dependent on V.

Since $f_\rho^V(x) \in [-1, 1]$, we can improve the error estimates by replacing values of f by projections onto $[-1, 1]$. The idea of the following projection operator was firstly introduced for this purpose in [29].

Definition 6. The projection operator π is defined on the space of measurable functions $f : X \to \mathbb{R}$ as

$$\pi(f)(x) = \begin{cases} 1, & \text{if } f(x) > 1, \\ -1, & \text{if } f(x) < -1, \\ f(x), & \text{if } -1 \le f(x) \le 1. \end{cases} \tag{15}$$

The definition of classifying loss implies that $V(y\pi(f)(x)) \le V(yf(x))$, so

$$\mathscr{E}(\pi(f)) \le \mathscr{E}(f), \qquad \mathscr{E}_{\mathbf{z}}(\pi(f)) \le \mathscr{E}_{\mathbf{z}}(f). \tag{16}$$

It is trivial that $\operatorname{sgn}(\pi(f)) = \operatorname{sgn}(f)$. By Proposition 5,

$$\mathscr{R}(\operatorname{sgn}(f)) - \mathscr{R}(f_c) \le \begin{cases} \mathscr{E}(\pi(f)) - \mathscr{E}(f_c) & \text{if } V(t) = (1-t)_+, \\ c_V \sqrt{\mathscr{E}(\pi(f)) - \mathscr{E}\left(f_\rho^V\right)} & \text{if } V''(0) \ge 0. \end{cases} \tag{17}$$

So it is sufficient for us to bound (13) by means of $\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}(f_\rho^V)$, which in turn can be estimated by an error decomposition technique. However, there are essential differences between algorithm (8) and (10). For example, the hypothesis space $\mathscr{H}_{K,\mathbf{z}}$ and the regularizer $\Omega_{\mathbf{z}}(f)$ in (10) are dependent on samples $\mathbf{z}$. This causes that the standard error analysis methods of (8) (see, e.g., [8, 12, 13, 30]) cannot be applied to (10) any more. This difficulty was overcome in [28] by introducing a modified error decomposition with an extra hypothesis error term. In this paper we apply the same underlying idea to classification scheme (10). To this end, we need to consider a Banach space containing all of the possible hypothesis space $\mathscr{H}_{K,\mathbf{z}}$.

Definition 7. The Banach space $\mathscr{H}_0$ is defined as the function set on X containing all functions of the form

$$f = \sum_{j=1}^{\infty} a_j K_{u_j}, \quad \{a_j\}_{j=1}^{\infty} \in l^1, \ \{u_j\}_{j=1}^{\infty} \subset X, \tag{18}$$

with the norm

$$\|f\| := \inf\left\{ \sum_{j=1}^{\infty} |a_j| : f = \sum_{j=1}^{\infty} a_j K_{u_j} \right\}. \tag{19}$$

Obviously,

$$\mathscr{H}_{K,\mathbf{z}} \subset \mathscr{H}_0, \quad \forall \mathbf{z} \in Z^m. \tag{20}$$

By the continuity of K and compactness of X, we have

$$\kappa := \sup_{x \in X} K(x, x) < \infty. \tag{21}$$

It implies that $\mathscr{H}_0$ is a subset of the continuous function space $C(X)$, and

$$\|f\|_\infty \le \kappa \|f\|, \quad \forall f \in \mathscr{H}_0. \tag{22}$$

To formulate the error decomposition for scheme (10), we introduce a regularization function as

$$f_\lambda := \arg\min_{f \in \mathscr{H}_0} \{\mathscr{E}(f) + \lambda \|f\|\}. \tag{23}$$

Proposition 8. *Let $f_{\mathbf{z},\lambda}$ be defined in* (10)*, $\lambda > 0$; then*

$$\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}\left(f_\rho^V\right) + \lambda \|f_{\mathbf{z},\lambda}\| \le S(\mathbf{z}, \lambda) + P(\mathbf{z}, \lambda) + D(\lambda). \tag{24}$$

Here

$$\begin{aligned} S(\mathbf{z}, \lambda) &:= \{\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))\} \\ &\quad + \{\mathscr{E}_{\mathbf{z}}(f_\lambda) - \mathscr{E}(f_\lambda)\}, \\ P(\mathbf{z}, \lambda) &:= \{\mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda})\} \\ &\quad - \{\mathscr{E}_{\mathbf{z}}(f_\lambda) + \lambda \|f_\lambda\|\}, \end{aligned} \tag{25}$$

$$\begin{aligned} D(\lambda) &:= \left\{\mathscr{E}(f_\lambda) - \mathscr{E}\left(f_\rho^V\right) + \lambda \|f_\lambda\|\right\} \\ &= \inf_{f \in \mathscr{H}_0} \left\{\mathscr{E}(f) - \mathscr{E}\left(f_\rho^V\right) + \lambda \|f\|\right\}. \end{aligned} \tag{26}$$

Proof.

$$\begin{aligned} &\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}\left(f_\rho^V\right) + \lambda \|f_{\mathbf{z},\lambda}\| \\ &\le \mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}\left(f_\rho^V\right) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \\ &= \{\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))\} \\ &\quad + \{\mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda})\} \\ &\quad + \{\mathscr{E}_{\mathbf{z}}(f_\lambda) - \mathscr{E}(f_\lambda)\} - \{\mathscr{E}_{\mathbf{z}}(f_\lambda) + \lambda \|f_\lambda\|\} \\ &\quad + \left\{\mathscr{E}(f_\lambda) - \mathscr{E}\left(f_\rho^V\right) + \lambda \|f_\lambda\|\right\} \\ &= S(\mathbf{z}, \lambda) + P(\mathbf{z}, \lambda) + D(\lambda). \end{aligned} \tag{27}$$

□

$P(\mathbf{z}, \lambda)$ and $S(\mathbf{z}, \lambda)$ are called hypothesis error and sample error, and they will be estimated, respectively, in next two sections. $D(\lambda)$ is independent of samples and usually called approximation error, and it characterizes the approximation ability of the function space $\mathscr{H}_0$ with respect to target function f_ρ^V. We will assume that, for some constants $0 < \beta \le 1$ and $c_\beta > 0$,

$$D(\lambda) \le c_\beta \lambda^\beta, \quad \forall \lambda > 0. \tag{28}$$

3. Estimating the Hypothesis Error

In this section we bound the hypothesis error $P(\mathbf{z}, \lambda)$ by a technique of scattered data interpolation which was firstly used in kernel regression context in [25]. To this end, we need some assumptions on input space X, margin distribution ρ_X, and kernel K. Denote $|\cdot|_2$ the Euclidean norm in $\mathbb{R}^n$.

Definition 9. A subset X of $\mathbb{R}^n$ is said to satisfy an interior cone condition if there exist an angle $\vartheta \in (0, \pi/2)$, a radius $R > 0$, and a unit vector $\zeta(x)$ for every $x \in X$ such that the cone

$$C(x, \zeta(x), \vartheta, R) = \left\{x + tw : w \in \mathbb{R}^n, |w|_2 = 1, w^T \zeta(x) \geq \cos\vartheta, t \in [0, R]\right\} \tag{29}$$

is contained in X.

Definition 10. The margin distribution ρ_X is said to satisfy condition L_τ with $0 < \tau < \infty$ if for some $c_\tau > 0$

$$\rho_X\left(\{u \in X : |u - x|_2 < r\}\right) \geq c_\tau r^\tau, \quad \forall x \in X,\ 0 < r \leq 1. \tag{30}$$

Recall that, for $s \in \mathbb{N}$, the space $C^s(X)$ consists of functions f whose partial derivative $D^d f$ is continuous for every $d = (d_1, d_2, \ldots, d_n) \in \mathbb{N}^n$ with $|d| := d_1 + \cdots + d_n \leq s$, and $\|f\|_{C^s} := \sum_{|d|\leq s} \|D^d f\|_\infty$. Throughout the paper we assume the kernel $K \in C^s(X \times X)$ with $s > 0$.

Definition 11. a set $\{x_1, x_2, \ldots, x_m\} \subset X$ is said to be Δ-dense if for any $x \in X$ there exists some $1 \leq i \leq m$ such that $|x - x_i|_2 < \Delta$.

The following lemma derived from [31] describes a local polynomial reproduction and it is the key point to bound the hypothesis error.

Lemma 12. *Suppose that $X \subset \mathbb{R}^n$ is compact and satisfies an interior cone condition with some radius $R > 0$ and angle $\vartheta \in (0, \pi/2)$. Fix $s \in \mathbb{N}$ with $s \geq 2$. Assume that the point set $\{x_1, x_2, \ldots, x_m\} \subset X$ is Δ-dense with $\Delta \leq R/c_0$ for some constant c_0 depending on X and s; then, for any $u \in X$, there exist real number $b_i(u)$, $i = 1, 2, \ldots, m$, satisfying that*

(1) $\sum_{i=1}^m b_i(u) p(x_i) = p(u)$, *$p(\cdot)$ is any polynomial of degree at most s on $\mathbb{R}^n$,*

(2) $\sum_{i=1}^m |b_i(u)| \leq 2$,

(3) $b_i(u) = 0$ *for those u satisfying $|u - x_i|_2 > c_0\Delta$.*

Proposition 13. *Let V be a classifying loss satisfying* (3)*, $K \in C^s(X \times X)$ for some $s \in \mathbb{N}$, $s \geq 2$. If X satisfies the conditions in Lemma 12 and $\{x_1, x_2, \ldots, x_m\}$ is Δ-dense in X with $\Delta \leq R/c_0$, then*

$$P(\mathbf{z}, \lambda) \leq D(\lambda) + 2c_\theta \|K\|_{C^s}\left(2^{\theta-1} + (4\kappa)^{\theta-1}\right) \times \left(\left(\frac{D(\lambda)}{\lambda}\right)^\theta + \frac{D(\lambda)}{\lambda}\right) c_0^s \Delta^s. \tag{31}$$

Proof. We know from (26) that $\|f_\lambda\| \leq D(\lambda)/\lambda$. So for any $\eta > 0$, f_λ can be written as $f_\lambda = \sum_{j=1}^\infty \beta_j K_{u_j}$ with $u_j \in X$ and

$$\|f_\lambda\| \leq \sum_{j=1}^\infty |\beta_j| < \|f_\lambda\| + \eta \leq \frac{D(\lambda)}{\lambda} + \eta. \tag{32}$$

At the same time, there exists some $N \in \mathbb{N}$ such that $\sum_{j=N}^\infty |\beta_j| < \eta$, and thus

$$\left\|\sum_{j=1}^N \beta_j K_{u_j} - f_\lambda\right\|_\infty \leq \kappa \sum_{j=N}^\infty |\beta_j| \leq \kappa\eta. \tag{33}$$

Fix $x \in X$ and $j \in \{1, 2, \ldots, N\}$, and we can take p_x as the Taylor polynomial of K_x of degree $s - 1$ at u_j. Then by Lemma 12, there exists $\{b_i(u_j)\}_{i=1}^m \in \mathbb{R}^m$ such that

$$\sum_{i \in I(u_j)} b_i(u_j) p_x(x_i) = p_x(u_j) = K_x(u_j), \quad \sum_{i \in I(u_j)} |b_i(u_j)| \leq 2, \tag{34}$$

where $I(u_j) = \{i \in \{1, 2, \ldots, m\} : |x_i - u_j|_2 \leq c_0\Delta\}$. Moreover,

$$\max_{i \in I(u_j)} |K_x(x_i) - p_x(x_i)| \leq \|K\|_{C^s}(c_0\Delta)^s. \tag{35}$$

It follows from (34) that

$$\begin{aligned} &\left|K_{u_j}(x) - \sum_{i \in I(u_j)} b_i(u_j) K_{x_i}(x)\right| \\ &= \left|K_x(u_j) - \sum_{i \in I(u_j)} b_i(u_j) K_x(x_i)\right| \\ &= \left|\sum_{i \in I(u_j)} b_i(u_j)\left(p_x(x_i) - K_x(x_i)\right)\right| \\ &\leq 2\|K\|_{C^s}(c_0\Delta)^s. \end{aligned} \tag{36}$$

The above bound holds for every $x \in X$ and $j \in \{1, 2, \ldots, N\}$, so

$$\left\|\sum_{j=1}^N \beta_j\left(K_{u_j} - \sum_{i \in I(u_j)} b_i(u_j) K_{x_i}\right)\right\|_\infty \leq 2\|K\|_{C^s}(c_0\Delta)^s \sum_{j=1}^N |\beta_j|. \tag{37}$$

This together with (33) and (32) implies

$$\left\|\sum_{j=1}^N \beta_j \sum_{i \in I(u_j)} b_i(u_j) K_{x_i} - f_\lambda\right\|_\infty \leq \kappa\eta + 2\|K\|_{C^s}(c_0\Delta)^s\left(\frac{D(\lambda)}{\lambda} + \eta\right). \tag{38}$$

Denote $f_0 = \sum_{j=1}^{N} \beta_j \sum_{i\in I(u_j)} b_i(u_j) K_{x_i} \in \mathcal{H}_{K,\mathbf{z}}$; we get from (16), (10), and (32) that

$$\begin{aligned}\mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) &\le \mathscr{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \\ &\le \mathscr{E}_{\mathbf{z}}(f_0) + \lambda\Omega_{\mathbf{z}}(f_0) \\ &\le \mathscr{E}_{\mathbf{z}}(f_0) + 2\lambda(\|f_\lambda\| + \eta).\end{aligned} \tag{39}$$

Since V is convex and satisfies (3), we have, for any $t_1, t_2 \in \mathbb{R}$,

$$\begin{aligned}|V(t_1) - V(t_2)| &\le \max\left\{\left|V'_{\pm}(t_1)\right|, \left|V'_{\pm}(t_2)\right|\right\}|t_1 - t_2| \\ &\le c_\theta(1 + \max\{|t_1|, |t_2|\})^{\theta-1}|t_1 - t_2|.\end{aligned} \tag{40}$$

This in connect with (22), (32), and (38) yields

$$\begin{aligned}&|\mathscr{E}_{\mathbf{z}}(f_0) - \mathscr{E}_{\mathbf{z}}(f_\lambda)| \\ &\quad\le c_\theta\left(1 + 2\kappa\left(\frac{D(\lambda)}{\lambda} + \eta\right)\right)^{\theta-1}\|f_0 - f_\lambda\|_\infty \\ &\quad\le c_\theta\left(1 + 2\kappa\left(\frac{D(\lambda)}{\lambda} + \eta\right)\right)^{\theta-1} \\ &\qquad\times\left(\kappa\eta + 2\|K\|_{C^s}(c_0\Delta)^s\left(\frac{D(\lambda)}{\lambda} + \eta\right)\right).\end{aligned} \tag{41}$$

Therefore

$$\begin{aligned}&\mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \\ &\quad\le \mathscr{E}_{\mathbf{z}}(f_\lambda) + \lambda\|f_\lambda\| + D(\lambda) + 2\lambda\eta \\ &\qquad + c_\theta\left(1 + 2\kappa\left(\frac{D(\lambda)}{\lambda} + \eta\right)\right)^{\theta-1} \\ &\qquad\times\left(\kappa\eta + 2\|K\|_{C^s}(c_0\Delta)^s\left(\frac{D(\lambda)}{\lambda} + \eta\right)\right).\end{aligned} \tag{42}$$

Let $\eta \to 0$, and we then derive

$$\begin{aligned}P(\mathbf{z},\lambda) &\le D(\lambda) + 2c_\theta\|K\|_{C^s}\left(2^{\theta-1} + (4\kappa)^{\theta-1}\right) \\ &\quad\times\left(\left(\frac{D(\lambda)}{\lambda}\right)^\theta + \frac{D(\lambda)}{\lambda}\right)c_0^s\Delta^s.\end{aligned} \tag{43}$$

□

We can now bound $P(\mathbf{z},\lambda)$ by the following theorem.

Theorem 14. *Let V be a classifying loss satisfying* (3), *$K \in C^s(X\times X)$ for some $s \in \mathbb{N}, s \ge 2$. Suppose that X satisfies the conditions in Lemma 12 and ρ_X satisfies condition L_τ with some $\tau > 0$, and* (28) *is valid; then, for any $0 < \delta < 1$ and m satisfying*

$$m \ge C_1\left(\log\left(\frac{2}{\delta}\right) + \log(m+1)\right), \tag{44}$$

with confidence $1 - \delta/2$,

$$\begin{aligned}&P(\mathbf{z},\lambda) \\ &\le C_2\left(\lambda^\beta + \left(\lambda^{\beta-1} + \lambda^{(\beta-1)\theta}\right)\right. \\ &\qquad\left.\times\left(\frac{\log(2/\delta) + \log(m+1)}{m}\right)^{s/\tau}\right),\end{aligned} \tag{45}$$

where C_1, C_2 are some constants independent of λ, m, or δ.

Proof. Applying Lemma 3 in [21], we get that, with confidence $1 - \delta/2$, the point set $\{x_1, x_2, \ldots, x_m\}$ is $A((\log(2/\delta) + \log(m+1))/m)^{1/\tau}$-dense in X, where A is a constant depending on X, τ, and c_τ. Taking $\Delta = A((\log(2/\delta) + \log(m+1))/m)^{1/\tau}$, $C_1 = (c_0 A/R)^\tau$. If $m \ge C_1(\log(2/\delta) + \log(m+1))$, then we have $\Delta \le R/c_0$. So Proposition 13 ensures us that, with confidence $1 - \delta/2$,

$$\begin{aligned}&P(\mathbf{z},\lambda) \\ &\le c_\beta\lambda^\beta + 2c_\theta\|K\|_{C^s}\left(2^{\theta-1} + (4\kappa)^{\theta-1}\right) \\ &\times\left(c_\beta^\theta\lambda^{(\beta-1)\theta} + c_\beta\lambda^{(\beta-1)}\right)c_0^s A^s\left(\frac{\log(2/\delta) + \log(m+1)}{m}\right)^{s/\tau}.\end{aligned} \tag{46}$$

This proves the theorem by setting $C_2 := c_\beta + 2c_\theta\|K\|_{C^s}(2^{\theta-1} + (4\kappa)^{\theta-1})c_0^s A^s(c_\beta^\theta + c_\beta)$. □

4. Estimating the Sample Error

In this section we focus on the sample error, it is the major improvement we make in this paper for the error analysis of algorithm (10).

Definition 15. Let $\mathscr{F}$ be a class of functions on Z and $\mathbf{z} := \{z_i\}_{i=1}^m \in Z^m$. The l^2-metric $d_{2,\mathbf{z}}$ is defined on $\mathscr{F}$ by

$$d_{2,\mathbf{z}}(f,g) := \left\{\frac{1}{m}\sum_{i=1}^{m}|f(z_i) - g(z_i)|^2\right\}^{1/2}. \tag{47}$$

For every $\varepsilon > 0$, the covering number of $\mathscr{F}$ with respect to $d_{2,\mathbf{z}}$ is

$$\begin{aligned}&\mathcal{N}_{2,\mathbf{z}}(\mathscr{F},\varepsilon) \\ &:= \inf\left\{l \in \mathbb{N} : \exists\{f_i\}_{i=1}^l\right. \\ &\qquad\left.\text{such that } \mathscr{F} = \cup_{i=1}^l\{f \in \mathscr{F} : d_{2,\mathbf{z}}(f,f_i) \le \varepsilon\}\right\}.\end{aligned} \tag{48}$$

The function sets in our situation are balls of $\mathscr{H}_0$ in the form of $\mathscr{B}_R = \{f \in \mathscr{H}_0 : \|f\| \le R\}$. We need the l^2-empirical covering number of $\mathscr{B}_1$ defined as

$$\mathcal{N}_2(\mathscr{B}_1,\varepsilon) := \sup_{m\in\mathbb{N}}\sup_{\mathbf{x}\in X^m}\mathcal{N}_{2,\mathbf{x}}(\mathscr{B}_1,\varepsilon). \tag{49}$$

According to a bound for l^2-empirical covering number derived in [32], we know that if $K \in C^s(X \times X)$, then

$$\log \mathcal{N}_2\left(\mathcal{B}_1, \varepsilon\right) \leq c_p\left(\frac{1}{\varepsilon}\right)^p, \quad \forall \varepsilon > 0, \tag{50}$$

where c_p is a constant independent of $\varepsilon > 0$, and $p \in (0, 2)$ is a power index defined dy

$$p = \begin{cases} \dfrac{2n}{2n+s}, & \text{when } 0 < s \leq 1, \\ \dfrac{2n}{2n+2}, & \text{when } 1 < s \leq 1 + \dfrac{n}{2}, \\ \dfrac{n}{s}, & \text{when } s > 1 + \dfrac{n}{2}. \end{cases} \tag{51}$$

For a measurable function $f : Z \to \mathbb{R}$, denote $\mathbb{E}f := \int_Z f(z)d\rho$. The following definition is a variance-expectation condition for the pair (V, ρ), which is generally used to achieve tight bounds.

Definition 16. A variance power α of the pair (V, ρ) is a number in $[0, 1]$ such that for any $f : X \to [-1, 1]$, there exists some constant $c_\alpha > 0$ satisfying

$$\mathbb{E}\left[V\left(yf\left(x\right)\right) - V\left(yf_\rho^V\left(x\right)\right)\right]^2 \leq c_\alpha\left[\mathscr{E}\left(f\right) - \mathscr{E}\left(f_\rho^V\right)\right]^\alpha. \tag{52}$$

Remark 17. It is easy to see that (52) always holds for $\alpha = 0$ and $c_\alpha = V^2(-1)$. Larger α is possible when V has strong convexity or ρ satisfies some noise condition (see [1, 5]).

We are in a position to bound the sample error. Write $S(\mathbf{z}, \lambda)$ as

$$\begin{aligned} &S\left(\mathbf{z}, \lambda\right) \\ &= \left\{\left[\mathscr{E}\left(\pi\left(f_{\mathbf{z},\lambda}\right)\right) - \mathscr{E}\left(f_\rho^V\right)\right] - \left[\mathscr{E}_{\mathbf{z}}\left(\pi\left(f_{\mathbf{z},\lambda}\right)\right) - \mathscr{E}_{\mathbf{z}}\left(f_\rho^V\right)\right]\right\} \\ &\quad + \left\{\left[\mathscr{E}_{\mathbf{z}}\left(f_\lambda\right) - \mathscr{E}_{\mathbf{z}}\left(f_\rho^V\right)\right] - \left[\mathscr{E}\left(f_\lambda\right) - \mathscr{E}\left(f_\rho^V\right)\right]\right\} \\ &:= S_1\left(\mathbf{z}, \lambda\right) + S_2\left(\mathbf{z}, \lambda\right). \end{aligned} \tag{53}$$

We will first bound $S_2(\mathbf{z}, \lambda)$, and to this end we need the following one-side Bernstein inequality (see [33]).

Let ξ be a random variable on a probability space Z with mean $\mathbb{E}\xi = \mu$ and variance $\sigma^2(\xi) = \sigma^2$. If $|\xi - \mu| \leq B$ almost everywhere, then for all $\eta > 0$

$$\begin{aligned} &\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{1}{m}\sum_{i=1}^m \xi\left(z_i\right) - \mu \geq \eta\right\} \\ &\qquad \leq \exp\left\{-\frac{m\eta^2}{2\left(\sigma^2 + (1/3)B\eta\right)}\right\}. \end{aligned} \tag{54}$$

Proposition 18. *Suppose that classifying loss V satisfies* (3). *If* (28) *and* (52) *hold, then, for any $0 < \delta < 1$ with confidence $1 - \delta/4$,*

$$S_2\left(\mathbf{z}, \lambda\right) \leq C_3 \log\frac{8}{\delta}\left(\left(\frac{1}{m}\right)^{1/(2-\alpha)} + \lambda^\beta + \frac{\lambda^{(\beta-1)\theta}}{m}\right), \tag{55}$$

where C_3 is a constant independent of λ, m, or δ.

Proof. Denote $\xi_1 := V(yf_\lambda(x)) - V(y\pi(f_\lambda)(x))$, $\xi_2 := V(y\pi(f_\lambda)(x)) - V(yf_\rho^V(x))$. Then

$$S_2\left(\mathbf{z}, \lambda\right) = \left\{\frac{1}{m}\sum_{i=1}^m \xi_1\left(z_i\right) - \mathbb{E}\xi_1\right\} + \left\{\frac{1}{m}\sum_{i=1}^m \xi_2\left(z_i\right) - \mathbb{E}\xi_2\right\}. \tag{56}$$

By (22) and (26), we can see that

$$\left\|f_\lambda\right\|_\infty \leq \kappa\left\|f_\lambda\right\| \leq \kappa\frac{D\left(\lambda\right)}{\lambda}. \tag{57}$$

We may assume $|f_\lambda(x)| > 1$, since otherwise $\xi_1 = 0$. Then from (3), we can derive that

$$\begin{aligned} 0 \leq \xi_1 \leq V\left(yf_\lambda\left(x\right)\right) &\leq c_\theta 2^\theta\left(\kappa\frac{D\left(\lambda\right)}{\lambda}\right)^\theta, \\ \sigma^2\left(\xi_1\right) &\leq c_\theta 2^\theta\left(\kappa\frac{D\left(\lambda\right)}{\lambda}\right)^\theta \mathbb{E}\xi_1. \end{aligned} \tag{58}$$

Applying the one-side Bernstein inequality to ξ_1 we have, for any $t > 1$ with confidence $1 - e^{-t}$,

$$\begin{aligned} &\frac{1}{m}\sum_{i=1}^m \xi_1\left(z_i\right) - \mathbb{E}\xi_1 \\ &\leq \frac{2c_\theta 2^\theta\kappa^\theta t}{3m}\left(\frac{D\left(\lambda\right)}{\lambda}\right)^\theta + \sqrt{\frac{2c_\theta 2^\theta\kappa^\theta t}{m}\left(\frac{D\left(\lambda\right)}{\lambda}\right)^\theta \mathbb{E}\xi_1} \\ &\leq \frac{2c_\theta 2^\theta\kappa^\theta t}{3m}\left(\frac{D\left(\lambda\right)}{\lambda}\right)^\theta + \frac{c_\theta 2^\theta\kappa^\theta t}{2m}\left(\frac{D\left(\lambda\right)}{\lambda}\right)^\theta + \mathbb{E}\xi_1 \\ &= \frac{7c_\theta 2^\theta\kappa^\theta t}{6m}\left(\frac{D\left(\lambda\right)}{\lambda}\right)^\theta + \mathbb{E}\xi_1. \end{aligned} \tag{59}$$

On the other hand, both $\pi(f_\lambda)(x)$ and $f_\rho^V(x)$ are contained in $[-1, 1]$, and we know from (3) and (52) that

$$\left|\xi_2\right| \leq c_\theta 2^\theta, \qquad \sigma^2\left(\xi_2\right) \leq c_\alpha(\mathbb{E}\xi_2)^\alpha. \tag{60}$$

Applying the one-side Bernstein inequality again, we have, with confidence $1 - e^{-t}$,

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^m \xi_2\left(z_i\right) - \mathbb{E}\xi_2 &\leq \frac{4c_\theta 2^\theta t}{3m} + \sqrt{\frac{2tc_\alpha(\mathbb{E}\xi_2)^\alpha}{m}} \\ &\leq \frac{4c_\theta 2^\theta t}{3m} + \frac{\alpha}{2}\mathbb{E}\xi_2 + \left(1 - \frac{\alpha}{2}\right)\left(\frac{2c_\alpha t}{m}\right)^{1/(2-\alpha)} \\ &\leq \frac{4c_\theta 2^\theta t}{3m} + \left(\frac{2c_\alpha t}{m}\right)^{1/(2-\alpha)} + \mathbb{E}\xi_2, \end{aligned} \tag{61}$$

where in the second inequality we have used the elementary inequality

$$\begin{aligned} &\frac{1}{\varrho} + \frac{1}{\varrho^*} = 1, \\ &\text{with } \varrho, \varrho^* > 1 \Longrightarrow ab \leq \frac{1}{\varrho}a^\varrho + \frac{1}{\varrho^*}b^{\varrho^*}, \quad \forall a, b > 0. \end{aligned} \tag{62}$$

Since $\mathbb{E}\xi_1 + \mathbb{E}\xi_2 = \mathscr{E}(f_\lambda) - \mathscr{E}(f_\rho^V) \le D(\lambda)$, combining the estimates above, we can get that, under assumption (28) with confidence $1 - 2e^{-t}$,

$$S_2(\mathbf{z}, \lambda) \le \frac{7c_\theta 2^\theta \kappa^\theta c_\beta^\theta \lambda^{(\beta-1)^\theta} t}{6m} + \frac{4c_\theta 2^\theta t}{3m} + \left(\frac{2c_\alpha t}{m}\right)^{1/(2-\alpha)} + c_\beta \lambda^\beta. \tag{63}$$

Then we prove the proposition by setting $t = \log(8/\delta)$, and

$$C_3 = \frac{7c_\theta 2^\theta \kappa^\theta c_\beta^\theta}{6} + \frac{4c_\theta 2^\theta}{3} + (2c_\alpha)^{1/(2-\alpha)} + c_\beta. \tag{64}$$

□

It is more difficult to bound $S_1(\mathbf{z}, \lambda)$, because it involves the samples $\mathbf{z}$ and thus runs over a set of functions. To get a better error estimation, An iteration technique is often used to shrink the radius of the ball containing $f_{\mathbf{z},\lambda}$ (see, e.g., [5, 12, 30, 32]); however this process is rather tough and complicated. In this paper We succeed to avoid the prolix iteration by considering the following reweighted empirical process

$$\left\{\omega_r(f)\left[\left(\mathscr{E}(\pi(f)) - \mathscr{E}\left(f_\rho^V\right)\right) - \left(\mathscr{E}_{\mathbf{z}}(\pi(f)) - \mathscr{E}_{\mathbf{z}}\left(f_\rho^V\right)\right)\right] : f \in \mathscr{H}_0\right\}. \tag{65}$$

Here $\omega_r(f) = (r + \omega(f))^{-1}$ for a threshold $r > 0$ and $\omega(f) = \mathscr{E}(\pi(f)) - \mathscr{E}(f_\rho^V) + \lambda\|f\|$. Different from the classical weight function, $\omega(f)$ contains the regularization term $\lambda\|f\|$ and thus makes it possible to control the variances and $\|f\|$ by the threshold r simultaneously.

The following concentration inequality is a scaled version of Theorem 2.3 in [34], where the case $B = 1$ is given.

Lemma 19. *Assume that $z_1, \ldots, z_m$ are identically distributed according to ρ. Let $\mathscr{F}$ be a countable set of measurable functions from Z to $[-B, B]$ and assume that all functions g in $\mathscr{F}$ satisfy $\mathbb{E}g = 0$, $\sigma^2(g) \le \sigma^2$ for some positive real number σ. Denote*

$$\xi = \sup_{g\in\mathscr{F}} \left|\frac{1}{m}\sum_{i=1}^m g(z_i)\right|. \tag{66}$$

Then, for all $t > 0$, one has

$$Prob\left\{\xi \ge \mathbb{E}\xi + \sqrt{\frac{2t\left(\sigma^2 + 2B\mathbb{E}\xi\right)}{m}} + \frac{Bt}{3m}\right\} \le e^{-t}. \tag{67}$$

This lemma allows us to take care of the deviation of the supremum of a empirical process with respect to its expectation.

Proposition 20. *Let $r \ge 0$ and*

$$\Phi_r := \sup_{f\in\mathscr{H}_0} \omega_r(f)\left|\left[\mathscr{E}(\pi(f)) - \mathscr{E}\left(f_\rho^V\right)\right] - \left[\mathscr{E}_{\mathbf{z}}(\pi(f)) - \mathscr{E}_{\mathbf{z}}\left(f_\rho^V\right)\right]\right|. \tag{68}$$

If (3) and (52) are satisfied, then, for any $t > 0$ with confidence $1 - e^{-t}$,

$$\Phi_r \le 2\mathbb{E}\Phi_r + \sqrt{\frac{2tc_\alpha r^{\alpha-2}}{m}} + \frac{8c_\theta 2^\theta t}{3mr}. \tag{69}$$

Proof. Before presenting the proof, let us first introduce some additional notations:

$$g_f(z) := V\left(y\pi(f)(x)\right) - V\left(yf_\rho^V(x)\right),$$
$$h_f^r(z) := \omega_r(f)\left[\mathbb{E}g_f - g_f(z)\right] = \frac{1}{\mathbb{E}g_f + \lambda\|f\| + r}\left[\mathbb{E}g_f - g_f(z)\right]. \tag{70}$$

Then

$$\Phi_r = \sup_{f\in\mathscr{H}_0} \frac{1}{m}\left|\sum_{i=1}^m h_f^r(z_i)\right|. \tag{71}$$

By (3) and (52), we can see that

$$\left|h_f^r(z)\right| \le \frac{2c_\theta 2^\theta}{r},$$
$$\sigma^2\left(h_f^r\right) \le \frac{\mathbb{E}\left(g_f\right)^2}{\left(\mathbb{E}g_f + r\right)^2} \le \frac{c_\alpha\left(\mathbb{E}g_f\right)^\alpha}{\left((2/\alpha)\,\mathbb{E}g_f\right)^\alpha \left((2/(2-\alpha))\,r\right)^{2-\alpha}} \le c_\alpha r^{\alpha-2}. \tag{72}$$

Here in the second inequality of (72), we have used the elementary inequality (62) again with $\varrho = 2/\alpha$, $a = (\varrho\mathbb{E}g_f)^{1/\varrho}$ and $b = (\varrho^* r)^{1/\varrho^*}$. So applying Lemma 19 to Φ_r, we get with confidence $1 - e^{-t}$

$$\Phi_r \le \mathbb{E}\Phi_r + \sqrt{\frac{2t\left(c_\alpha r^{\alpha-2} + \left(4c_\theta 2^\theta \mathbb{E}\Phi_r / r\right)\right)}{m}} + \frac{2c_\theta 2^\theta t}{3mr}$$
$$\le \mathbb{E}\Phi_r + \sqrt{\frac{2tc_\alpha r^{\alpha-2}}{m}} + \sqrt{\frac{8tc_\theta 2^\theta \mathbb{E}\Phi_r}{mr}} + \frac{2c_\theta 2^\theta t}{3mr} \tag{73}$$
$$\le 2\mathbb{E}\Phi_r + \sqrt{\frac{2tc_\alpha r^{\alpha-2}}{m}} + \frac{8c_\theta 2^\theta t}{3mr}.$$

□

So we can bound Φ_r through bounding its expectation. To this end, we need some preparations.

Definition 21. Let (Z, ρ) be a probability space, and $\mathscr{F}$ is a class of measurable functions from Z to $\mathbb{R}$. Set $\{z_i\}_{i=1}^m$ to be m independent random variables distributed according to ρ and $\{\epsilon_i\}_{i=1}^m$ to be m independent Rademacher random variables.

Then for $\eta > 0$ the local Rademacher average of $\mathscr{F}$ is defined by

$$\mathrm{Rad}\left(\mathscr{F}, m, \eta\right) := \mathbb{E} \sup_{\substack{f \in \mathscr{F} \\ \mathbb{E} f^2 \le \eta}} \left| \frac{1}{m} \sum_{i=1}^{m} \epsilon_i f\left(z_i\right) \right|. \tag{74}$$

The following lemma was given in [35].

Lemma 22. *Let $\mathscr{F}$ be a class of measurable functions from Z to $[-B, B]$. If for some $a > 0$ and $p \in (0, 2)$*

$$\sup_{m \in \mathbb{N}} \sup_{\mathbf{z} \in Z^m} \log \mathscr{N}_{2,\mathbf{z}}\left(\mathscr{F}, \varepsilon\right) \le a \varepsilon^{-p}, \quad \forall \varepsilon > 0, \tag{75}$$

then there exists a constant c_p' depending only on p such that

$$\begin{aligned} &Rad\left(\mathscr{F}, m, \eta\right) \\ &\le c_p' \max\left\{ \eta^{1/2 - p/4} \left(\frac{a}{m}\right)^{1/2}, B^{(2-p)/(2+p)} \left(\frac{a}{m}\right)^{2/(2+p)} \right\}. \end{aligned} \tag{76}$$

Lemma 23. *Let z be a nonnegative stochastic process indexed by T and $\omega(t)$ a nonnegative nonrandom function defined on T. Define $r^* = \inf_{t \in T} \omega(t)$. Let $\psi : [r^*, +\infty) \to [0, +\infty)$ be a function such that $\psi(4r) \le 4^{\nu} \psi(r)$ for some $0 < \nu < 1$, and*

$$\mathbb{E}\left[\sup_{\substack{t \in T \\ \omega(t) \le r}} z\left(t\right) \right] \le \psi\left(r\right), \quad \forall r \ge r^*. \tag{77}$$

Then, for any $r \ge r^*$, we have

$$\mathbb{E}\left[\sup_{t \in T} \frac{z\left(t\right)}{\omega\left(t\right) + r} \right] \le \left(1 + \frac{4^{\nu}}{1 - 4^{\nu - 1}}\right) \frac{\psi\left(r\right)}{r}. \tag{78}$$

Proof. For $r \ge r^*$, we obtain by a standard peeling approach

$$\sup_{t \in T} \frac{z\left(t\right)}{\omega\left(t\right) + r} \le \sup_{\substack{t \in T \\ \omega(t) \le r}} \frac{z\left(t\right)}{r} + \sum_{i=0}^{\infty} \sup_{\substack{t \in T \\ \omega(t) \in \left(4^i r, 4^{i+1} r\right]}} \frac{z\left(t\right)}{4^i r + r}. \tag{79}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[\sup_{t \in T} \frac{z\left(t\right)}{\omega\left(t\right) + r} \right] &\le \frac{\psi\left(r\right)}{r} + \sum_{i=0}^{\infty} \frac{\psi\left(4^{i+1} r\right)}{\left(1 + 4^i\right) r} \\ &\le \frac{\psi\left(r\right)}{r} + \frac{\psi\left(r\right)}{r} \sum_{i=0}^{\infty} \frac{4^{(i+1)\nu}}{1 + 4^i} \\ &\le \frac{\psi\left(r\right)}{r} \left(1 + \sum_{i=0}^{\infty} 4^{(\nu - 1) i + \nu}\right) \\ &= \left(1 + \frac{4^{\nu}}{1 - 4^{\nu - 1}}\right) \frac{\psi\left(r\right)}{r}. \end{aligned} \tag{80}$$

□

Now we can give a bound of Φ_r.

Proposition 24. *Let V be a classifying loss that satisfies* (3)*, $K \in C^s(X \times X)$ with $s > 0$, and $p \in (0, 2)$ given by* (51)*. Under assumption* (52)*, for any $t > 0$ and $r > 0$ satisfying*

$$\begin{aligned} r \ge \max \Bigg\{ &\left[C_4 \left(\frac{1}{m \lambda^p} \right)^{1/2} \right]^{4/(2-p)(2-\alpha)}, \\ &\left[C_4 \left(\frac{1}{m \lambda^p} \right)^{2/(2+p)} \right]^{(2+p)/(2-p)}, \\ &\left(\frac{32 t c_\alpha}{m} \right)^{1/(2-\alpha)}, \frac{c_\theta 2^{\theta+5} t}{3m}, D\left(\lambda\right) \Bigg\}, \end{aligned} \tag{81}$$

where $C_4 := 16(1 + (4^{\nu}/(1 - 4^{\nu-1})))c_p' \max\{c_\alpha^{(2-p)/4} (c_p (c_\theta 2^{\theta-1})^p)^{1/2}, (c_\theta 2^\theta)^{(2-p)/(2+p)} (c_p (c_\theta 2^{\theta-1})^p)^{2/(2+p)}\}$ *with* $\nu = \max\{((2-p)\alpha/4) + p/2, 2p/(2+p)\}$*. We have, with confidence* $1 - e^{-t}$*,*

$$\Phi_r \le \frac{3}{4}. \tag{82}$$

Proof. Using the notations in the proof of Proposition 20, the weight function

$$\begin{aligned} \omega\left(f\right) &= \mathbb{E} g_f + \lambda \left\| f \right\|, \\ r^* &:= \inf_{f \in \mathscr{H}_0} \omega\left(f\right) = D\left(\lambda\right). \end{aligned} \tag{83}$$

Hence $r \ge r^*$, and

$$\begin{aligned} &\mathbb{E}\left[\sup_{\substack{f \in \mathscr{H}_0 \\ \omega(f) \le r}} \left| \mathbb{E} g_f - \frac{1}{m} \sum_{i=1}^{m} g_f\left(z_i\right) \right| \right] \\ &\le \mathbb{E}\left[\sup_{\substack{f \in \mathscr{B}_{r/\lambda} \\ \mathbb{E} g_f \le r}} \left| \mathbb{E} g_f - \frac{1}{m} \sum_{i=1}^{m} g_f\left(z_i\right) \right| \right]. \end{aligned} \tag{84}$$

Denote $\mathscr{G} := \{g_f : f \in \mathscr{B}_{r/\lambda}\}$. Standard symmetrization argument (see Lemma 2.3.1 of [36]) and Assumption (52) then yield

$$\mathbb{E}\left[\sup_{\substack{f \in \mathscr{B}_{r/\lambda} \\ \mathbb{E} g_f \le r}} \left| \mathbb{E} g_f - \frac{1}{m} \sum_{i=1}^{m} g_f\left(z_i\right) \right| \right] \le 2 \mathrm{Rad}\left(\mathscr{G}, m, c_\alpha r^\alpha\right). \tag{85}$$

By (3) we know $|g_f(z)| \le c_\theta 2^\theta$, and

$$\begin{aligned} \left| g_{f_1}\left(z\right) - g_{f_2}\left(z\right) \right| &= \left| V\left(y \pi\left(f_1\right)\left(x\right)\right) - V\left(y \pi\left(f_2\right)\left(x\right)\right) \right| \\ &\le c_\theta 2^{\theta - 1} \left| \pi\left(f_1\right)\left(x\right) - \pi\left(f_2\right)\left(x\right) \right| \\ &\le c_\theta 2^{\theta - 1} \left| f_1\left(x\right) - f_2\left(x\right) \right|, \end{aligned} \tag{86}$$

for any $f_1, f_2 \in \mathscr{B}_{r/\lambda}$, so

$$\mathcal{N}_{2,\mathbf{z}}(\mathcal{G},\varepsilon) \le \mathcal{N}_{2,\mathbf{x}}\left(\mathscr{B}_{r/\lambda}, \frac{\varepsilon}{c_\theta 2^{\theta-1}}\right) = \mathcal{N}_{2,\mathbf{x}}\left(\mathscr{B}_1, \frac{\varepsilon\lambda}{c_\theta 2^{\theta-1} r}\right). \tag{87}$$

This together with (50) implies that

$$\sup_{m\in\mathbb{N}}\sup_{\mathbf{z}\in Z^m} \log \mathcal{N}_{2,\mathbf{z}}(\mathcal{G},\varepsilon) \le c_p\left(c_\theta 2^{\theta-1}\right)^p \left(\frac{r}{\lambda}\right)^p \varepsilon^{-p}. \tag{88}$$

By Lemma 22, we obtain

$$\begin{aligned}
&\mathrm{Rad}\left(\mathcal{G}, m, c_\alpha r^\alpha\right)\\
&\le c_p' \max\left\{ c_\alpha^{1/2-p/4} r^{\alpha(1/2-p/4)} \left(\frac{c_p (c_\theta 2^{\theta-1})^p}{m}\left(\frac{r}{\lambda}\right)^p\right)^{1/2},\right.\\
&\qquad \left.\left(c_\theta 2^\theta\right)^{(2-p)/(2+p)} \left(\frac{c_p\left(c_\theta 2^{\theta-1}\right)^p}{m}\left(\frac{r}{\lambda}\right)^p\right)^{2/(2+p)}\right\}\\
&:= \frac{\psi(r)}{2}.
\end{aligned} \tag{89}$$

Setting $\nu = \max\{((2-p)\alpha/4) + p/2, 2p/(2+p)\} \in (0,1)$, it is easy to see $\psi(4r) \le 4^\nu \psi(r)$. So Lemma 23 tells us

$$\begin{aligned}
\mathbb{E}\Phi_r &= \mathbb{E}\left[\sup_{f\in\mathscr{H}_0} \frac{\left|\mathbb{E}g_f - (1/m)\sum_{i=1}^m g_f(z_i)\right|}{\omega(f) + r}\right]\\
&\le \left(1 + \frac{4^\nu}{1-4^{\nu-1}}\right)\frac{\psi(r)}{r}.
\end{aligned} \tag{90}$$

By the choice of r, we can easily check that

$$\begin{aligned}
&\left(1 + \frac{4^\nu}{1-4^{\nu-1}}\right)\frac{\psi(r)}{r} \le \frac{1}{8}, \qquad \sqrt{\frac{2tc_\alpha r^{\alpha-2}}{m}} \le \frac{1}{4},\\
&\frac{8c_\theta 2^\theta t}{3mr} \le \frac{1}{4}.
\end{aligned} \tag{91}$$

Then the conclusion follows from Proposition 20. □

Corollary 25. *Under the condition of Proposition 24, if* (28) *is satisfied, then for any* $0 < \delta < 1$, *with confidence* $1 - \delta/4$, *there holds*

$$\begin{aligned}
&S_1(\mathbf{z},\lambda)\\
&\le \frac{3}{4}\left[\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}\left(f_\rho^V\right) + \lambda\left\|f_{\mathbf{z},\lambda}\right\|\right]\\
&\quad + C_5 \log\frac{4}{\delta}\left[\left(\frac{1}{m\lambda^p}\right)^{2/(2-p)(2-\alpha)} + \left(\frac{1}{m\lambda^p}\right)^{2/(2-p)}\right.\\
&\qquad \left.+\left(\frac{1}{m}\right)^{1/(2-\alpha)} + \lambda^\beta\right],
\end{aligned} \tag{92}$$

where C_5 *is a constant independent of* λ, m, *or* δ.

Proof. Taking

$$\begin{aligned}
r &= \left[C_4\left(\frac{1}{m\lambda^p}\right)^{1/2}\right]^{4/(2-p)(2-\alpha)}\\
&\quad + \left[C_4\left(\frac{1}{m\lambda^p}\right)^{2/(2+p)}\right]^{(2+p)/(2-p)}\\
&\quad + \left(\frac{32tc_\alpha}{m}\right)^{1/(2-\alpha)} + \frac{c_\theta 2^{\theta+5} t}{3m} + D(\lambda)
\end{aligned} \tag{93}$$

in Proposition 24, then, for any $t > 1$ with confidence $1 - e^{-t}$,

$$\begin{aligned}
&\omega_r(f_{\mathbf{z},\lambda})\left\{\left[\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}\left(f_\rho^V\right)\right]\right.\\
&\qquad \left. - \left[\mathscr{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}_{\mathbf{z}}\left(f_\rho^V\right)\right]\right\} \le \Phi_r \le \frac{3}{4}.
\end{aligned} \tag{94}$$

It follows that

$$\begin{aligned}
S_1(\mathbf{z},\lambda) &\le \frac{3}{4}\left(\omega(f_{\mathbf{z},\lambda}) + r\right)\\
&= \frac{3}{4}\left[\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}\left(f_\rho^V\right) + \lambda\left\|f_{\mathbf{z},\lambda}\right\|\right] + \frac{3}{4}r.
\end{aligned} \tag{95}$$

Then the corollary is proved by setting $t = \log(4/\delta)$, and

$$\begin{aligned}
C_5 := \frac{3}{4}\left(C_4^{4/(2-p)(2-\alpha)} + C_4^{(2+p)/(2-p)}\right.\\
\left. + (32c_\alpha)^{1/(2-\alpha)} + \frac{c_\theta 2^{\theta+5}}{3} + c_\beta\right).
\end{aligned} \tag{96}$$

□

5. Deriving Learning Rates

We may now present the main results by combining the results obtained in the previous two sections. The following theorem gives the bounds for the excess generalization error.

Theorem 26. *Let* V *be a classifying loss satisfying* (3), $K \in C^s(X\times X)$ *for some* $s \in \mathbb{N}$, $s \ge 2$, *and* $p \in (0,2)$ *given by* (51). *Suppose that* X *satisfies an interior cone conditions, and* ρ_X *satisfies condition* L_τ *with some* $\tau > 0$. *If* (28) *is valid, then for any* $0 < \delta < 1$ *and* m *satisfying* (44), *by taking* $\lambda = m^{-\gamma}$ *with* $\gamma = \min\{s/\tau(\beta + (1-\beta)\theta), 1/(\beta + (1-\beta)\theta), 2/(2-\alpha)(2-p)\beta + 2p\}$, *we have, with confidence* $1 - \delta$,

$$\begin{aligned}
&\mathscr{E}(\pi(f_{\mathbf{z},\lambda})) - \mathscr{E}\left(f_\rho^V\right)\\
&\le C\left(\log\frac{8}{\delta} + \log(m+1)\right)^{\max\{1,s/\tau\}}\left(\frac{1}{m}\right)^{\gamma\beta},
\end{aligned} \tag{97}$$

where C *is a constant independent of* m *or* δ.

Proof. Putting the estimates of Theorem 14, Proposition 18, and Corollary 25 into the error decomposition (24), we see that, with confidence $1-\delta$,

$$\begin{aligned}
&\mathscr{E}\left(\pi\left(f_{\mathbf{z},\lambda}\right)\right)-\mathscr{E}\left(f_\rho^V\right)+\lambda\left\|f_{\mathbf{z},\lambda}\right\| \\
&\le C_2\left(\lambda^\beta+\left(\lambda^{\beta-1}+\lambda^{(\beta-1)\theta}\right)\left(\frac{\log\left(2/\delta\right)+\log\left(m+1\right)}{m}\right)^{s/\tau}\right) \\
&\quad+C_3\log\frac{8}{\delta}\left(\left(\frac{1}{m}\right)^{1/(2-\alpha)}+\lambda^\beta+\frac{\lambda^{(\beta-1)\theta}}{m}\right) \\
&\quad+\frac{3}{4}\left[\mathscr{E}\left(\pi\left(f_{\mathbf{z},\lambda}\right)\right)-\mathscr{E}\left(f_\rho^V\right)+\lambda\left\|f_{\mathbf{z},\lambda}\right\|\right] \\
&\quad+C_5\log\frac{4}{\delta}\left[\left(\frac{1}{m\lambda^p}\right)^{2/(2-p)(2-\alpha)}+\left(\frac{1}{m\lambda^p}\right)^{2/(2-p)}\right. \\
&\qquad\left.+\left(\frac{1}{m}\right)^{1/(2-\alpha)}+\lambda^\beta\right].
\end{aligned} \tag{98}$$

Therefore, with the same confidence

$$\begin{aligned}
&\mathscr{E}\left(\pi\left(f_{\mathbf{z},\lambda}\right)\right)-\mathscr{E}\left(f_\rho^V\right) \\
&\le\mathscr{E}\left(\pi\left(f_{\mathbf{z},\lambda}\right)\right)-\mathscr{E}\left(f_\rho^V\right)+\lambda\left\|f_{\mathbf{z},\lambda}\right\| \\
&\le 4\left(C_2+C_3+C_5\right)\left(\log\frac{8}{\delta}+\log\left(m+1\right)\right)^{\max\{1,s/\tau\}} \\
&\quad\times\left[\lambda^\beta+\left(\frac{1}{m}\right)^{s/\tau}\left(\lambda^{\beta-1}+\lambda^{(\beta-1)\theta}\right)+\left(\frac{1}{m}\right)^{1/(2-\alpha)}\right. \\
&\qquad\left.+\frac{\lambda^{(\beta-1)\theta}}{m}+\left(\frac{1}{m\lambda^p}\right)^{2/(2-p)(2-\alpha)}+\left(\frac{1}{m\lambda^p}\right)^{2/(2-p)}\right].
\end{aligned} \tag{99}$$

By the choice of λ, we can easily check that

$$\begin{gathered}
\left(\frac{1}{m}\right)^{s/\tau}\lambda^{\beta-1}\le\left(\frac{1}{m}\right)^{s/\tau}\lambda^{(\beta-1)\theta}\le\lambda^\beta, \\
\left(\frac{1}{m}\right)^{1/(2-\alpha)}\le\lambda^\beta, \\
\frac{\lambda^{(\beta-1)\theta}}{m}\le\lambda^\beta,\qquad\left(\frac{1}{m\lambda^p}\right)^{2/(2-p)(2-\alpha)}\le\lambda^\beta, \\
\left(\frac{1}{m\lambda^p}\right)^{2/(2-p)}\le\lambda^\beta.
\end{gathered} \tag{100}$$

So our theorem follows by taking $C=28(C_2+C_3+C_5)$. □

Theorem 26 together with (17) allows us to give an explicit learning rate for misclassification error of scheme (10).

Corollary 27. *If the conditions in Theorem 26 are satisfied, then for any $0<\delta<1$, with confidence $1-\delta$, there holds*

$$\begin{aligned}
&\mathscr{R}\left(\operatorname{sgn}\left(f_{\mathbf{z},\lambda}\right)\right)-\mathscr{R}\left(f_c\right) \\
&\le\begin{cases}
C\left(\log\dfrac{8}{\delta}+\log\left(m+1\right)\right)^{\max\{1,s/\tau\}}\left(\dfrac{1}{m}\right)^{\gamma\beta} \\
\qquad\qquad\text{if } V(t)=(1-t)_+, \\
\widetilde{C}\left(\log\dfrac{8}{\delta}+\log\left(m+1\right)\right)^{\max\{1/2,s/2\tau\}}\left(\dfrac{1}{m}\right)^{\gamma\beta/2} \\
\qquad\qquad\text{if } V''(0)\ge 0,
\end{cases}
\end{aligned} \tag{101}$$

where $\widetilde{C}:=c_V\sqrt{C}$.

Remark 28. For the hinge loss $V_h(t)=(1-t)_+$, the increment exponent $\theta=1$. If $K\in C^\infty(X\times X)$, then one can take $s\to\infty$ and $p\to 0$. This is the case for polynomial kernel (see [13, 14]) or Gaussian kernel (see [5, 15]), usually used in practice. So Corollary 27 tells us the learning rate of the 1-norm SVM is $\mathcal{O}(m^{-\varsigma})$ with ς arbitrarily close to $\min\{\beta,1/(2-\alpha)\}$.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by National Natural Science Foundation of China under Grants 11171014 and 11072274, the Program for Innovative Research Team in UIBE.

References

[1] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

[2] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[3] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.

[4] Q. Wu and D. Zhou, "Analysis of support vector machine classification," *Journal of Computational Analysis and Applications*, vol. 8, no. 2, pp. 99–119, 2006.

[5] I. Steinwart and C. Scovel, "Fast rates for support vector machines using Gaussian kernels," *Annals of Statistics*, vol. 35, no. 2, pp. 575–607, 2007.

[6] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[7] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," *Annals of Statistics*, vol. 32, no. 1, pp. 56–134, 2004.

[8] D. R. Chen, Q. Wu, Y. M. Ying, and D. X. Zhou, "Support vector machine soft margin classifiers: error analysis," *Journal of Machine Learning Research*, vol. 5, pp. 1143–1175, 2004.

[9] Y. Lin, "Support vector machines and the Bayes rule in classification," *Data Mining and Knowledge Discovery*, vol. 6, no. 3, pp. 259–275, 2002.

[10] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[11] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[12] Q. Wu, Y. Ying, and D. Zhou, "Multi-kernel regularized classifiers," *Journal of Complexity*, vol. 23, no. 1, pp. 108–134, 2007.

[13] H. Z. Tong, D. R. Chen, and L. Z. Peng, "Learning rates for regularized classifiers using multivariate polynomial kernels," *Journal of Complexity*, vol. 24, no. 5-6, pp. 619–631, 2008.

[14] D. X. Zhou and K. Jetter, "Approximation with polynomial kernels and SVM classifiers," *Advances in Computational Mathematics*, vol. 25, no. 1–3, pp. 323–344, 2006.

[15] D. H. Xiang and D. X. Zhou, "Classification with Gassian and convex loss," *Journal of Machine Learning Research*, vol. 10, pp. 1447–1468, 2009.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B*, vol. 58, pp. 267–288, 1996.

[17] T. Zhang, "Some sharp performance bounds for least squares regression with L_1 regularization," *Annals of Statistics A*, vol. 37, no. 5, pp. 2109–2144, 2009.

[18] P. Zhao and B. Yu, "On model selection consistency of Lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[19] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[20] E. J. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, pp. 969–985, 2007.

[21] Q. W. Xiao and D. X. Zhou, "Learning by nonsymmetric kernels with data dependent spaces and l^1—regularizer," *Taiwanese Journal of Mathematics*, vol. 14, no. 5, pp. 1821–1836, 2010.

[22] H. Z. Tong, D. R. Chen, and F. H. Yang, "Least square regression with l^p—coefficient regularization," *Neural Computation*, vol. 22, no. 12, pp. 3221–3235, 2010.

[23] H. W. Sun and Q. Wu, "Least square regression with indefinite kernels and coefficient regularization," *Applied and Computational Harmonic Analysis*, vol. 30, no. 1, pp. 96–109, 2011.

[24] H. Z. Tong, D. R. Chen, and F. H. Yang, "Support vector machines regression with l^1—regularizer," *Journal of Approximation Theory*, vol. 164, no. 10, pp. 1331–1344, 2012.

[25] H. Y. Wang, Q. W. Xiao, and D. X. Zhou, "An approximation theory approach to learning with l^1 regularization," *Journal of Approximation Theory*, vol. 167, pp. 240–258, 2013.

[26] B. Tarigan and S. A. Van De Geer, "Classifiers of support vector machine type with l_1 complexity regularization," *Bernoulli*, vol. 12, no. 6, pp. 1045–1076, 2006.

[27] Q. Wu and D. Zhou, "SVM soft margin classifiers: linear programming versus quadratic programming," *Neural Computation*, vol. 17, no. 5, pp. 1160–1187, 2005.

[28] Q. Wu and D. Zhou, "Learning with sample dependent hypothesis spaces," *Computers and Mathematics with Applications*, vol. 56, no. 11, pp. 2896–2907, 2008.

[29] P. L. Bartlett, "The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 525–536, 1998.

[30] Q. Wu, Y. Ying, and D. Zhou, "Learning rates of least-square regularized regression," *Foundations of Computational Mathematics*, vol. 6, no. 2, pp. 171–192, 2006.

[31] H. Wendland, "Local polynomial reproduction and moving least squares approximation," *IMA Journal of Numerical Analysis*, vol. 21, no. 1, pp. 285–300, 2001.

[32] L. Shi, Y. Feng, and D. Zhou, "Concentration estimates for learning with l^1—regularizer and data dependent hypothesis spaces," *Applied and Computational Harmonic Analysis*, vol. 31, no. 2, pp. 286–302, 2011.

[33] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2002.

[34] O. Bousquet, "A Bennett concentration inequality and its application to suprema of empirical processes," *Comptes Rendus Mathematique*, vol. 334, no. 6, pp. 495–500, 2002.

[35] G. Blanchard, G. Lugosi, and N. Vayatis, "On the rate of convergence of regularized boosting classifiers," *Journal of Machine Learning Research*, vol. 4, no. 5, pp. 861–894, 2004.

[36] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, Springer, New York, NY, USA, 1996.